DOCUMENT RESUME

ED 227 616 EC 151 057

AUTHOR Elliott, Stephen; And Others
TITLE WISC-R-Test Stability and the Practice of Psychological Re-Evaluations.
PUB DATE Aug 82
NOTE 18p.; Paper presented at the Annual Convention of the American Psychological Association (90th, Washington, DC, August 23-27, 1982).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Academic Achievement; *Achievement Tests; *Disabilities; Elementary Secondary Education; *Ethnic Groups; Exceptional Child Research; *Intelligence Tests; Predictive Measurement; *Test Reliability

ABSTRACT

The study, involving 379 students (6-20 years old), investigated the 3-year performance stability of handicapped children from four racial groups (Anglo, Black, Mexican American, and Native American). Data were collected from the Wechsler Intelligence Scale for Children--Revised (WISC-R) and the Wide Range Achievement Test (WRAT). Results indicated that Anglo children's WISC-R test-retest reliabilities were higher than those for other ethnic groups; and retest WRAT subtest performances of all children were best predicted by initial WRAT subtest performances. However, overall stability of WRAT scores was only moderate. Results were discussed from an applied perspective in order to examine their impact on the practice of mandated special education re-evaluations. (Author/SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED227616

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

WISC-R Test Stability and the Practice of Psychological Re-evaluations

Stephen N. Elliott Wayne C. Piersel Joseph C. Witt
The University of Nebraska-Lincoln

Ed Argulewicz
Arizona State University

Paper presented at the
Annual Convention of the American Psychological Association
Washington, D. C.
August 1982

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Stephen N. Elliott
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

EC151057

WISC-R Stability and the Practice of Psychological Re-Evaluations

The stability of test scores has been a concern of many psychologists and educators. The unit of time across which one initially tests and then retests an individual, as well as the nature of the behavior being measured affects the stability of any measurement. Many longitudinal studies have been done which measure general intelligence and scholastic aptitude (Bayley, 1949; Harnquist, 1968; Honzik, MacFarlane, & Allen, 1948; Hopkins & Bracht, 1975; Zax, Cowen, Beach & Rappaport, 1972). These studies all demonstrated that stability is greater for shorter time periods than for longer time periods. In addition, the studies revealed increased stability with increased age.

After analyzing more than 20 longitudinal studies of intelligence, Bloom (1964) suggested that "in terms of intelligence measured at age 17, about 50% of the development takes place between conception and age 4, about 30% between ages 4 and 8, and about 20% between ages 8 and 17" (p. 88). From such a suggestion, one can deduce it would be important to occasionally (every 2 to 4 years) reassess children placed in special education to reaffirm that they continue to meet diagnostic placement criteria. In fact, since the passage of P.L. 94-142, the re-evaluation of children receiving special education services has become a mandated function of psychologists working in schools.

The concept of re-evaluating children in special education programs appears theoretically sound; it may well be the most frequently used, if not the only type of case follow-up employed by most school psychologists. The practices comprising the typical re-evalution, however, may not be sound (Elliott, Piersel, & Galvin, in press). That is, administering the same or a similar battery of norm-referenced tests of general abilities and achievement to a child approximately

every three years has not been demonstrated to provide psychologists or educators with useful information for designing educational programs. In addition, Vance, Blixt, Ellis, and Debell (1981) demonstrated that WISC-R Full Scale IQs were very stable over a 2-year interval for a sample of exceptional children. Thus, we believe it was necessary to investigate actual re-evaluation test data to gain more insight into the stability of test scores over time. Such information would shed light on the psychometric properities of various assessment instruments for different groups, and may also help determine the most cost-effective testing practices.

Statement of the Problem Investigated

In the present investigation, the practice of retesting exceptional children (specifically Anglo, Black, Mexican American, and Native American children) was examined to determine the stability of their performances on tests of general intelligence (i.e., Wechsler Intelligence Scale for Children-Revised) and achievement (i.e., Wide Range Achievement Test). Based on previous developmental and psychometric research, the investigators hypothesized that Anglo children's retest performance on the WISC-R would be more stable than the retest performances of Black, Mexican American, or Native American children.

Background Information

Two invesitgators, Martin (1979) and Thomas (1980), have examined the three year test-retest stability of exceptional children's performances on intelligence tests. Vance et al. (1981) also examined test-retest stability on an intelligence test; however, they utilized a two-year rather than a three-year interval. Martin's sample included black and white children classified as educable mentally retarded (EMR), behaviorally disordered (BD), and learning disabled (LD). A weakness of this study, however, was the use of a variety of intelligence measures. For example, some children were initially tested with the WISC and retested with the WISC-R, others were given the WISC first and the WAIS for the

re-evaluation, or the WPPSI first and the WISC-R for the re-evaluation. The three forms of the Wechsler Intelligence Scales are quite compatable; however, scores from one form cannot be directly equated with another form because the standardization samples were different and the content of each test varies. An additional problem with Martin's study was small sample sizes for the BD classification. Therefore, Martin's results must be interpreted with caution. In general, he found a wide range of variability in individual's IQs from the initial to re-evaluation. This was true across all three diagnostic groups. Generally, IQ decreased in children who were initially tested with the WISC and retested three years later with the WISC-R. Although black and white, male and female children were identified as subjects, no descriptive or analytic data were provided based on the independent variables of race or sex.

Thomas' (1980) investigation of children's three year test-retest performances on intelligence tests also utilized the WISC and the WISC-R. Her subjects included white and non-white (i.e., blacks and Mexican Americans) school children, but not in equal proportions. Hence, test results were not analyzed according to a racial or ethnic variable. Thomas classified her subjects into two groups for purposes of comparison. The first group had been given the WISC for both the initial and re-evaluation. The second group was given the WISC initially and the WISC-R for the re-evaluation. Thomas found that WISC and WISC-R scores did not change in a parallel fashion over a three-year period. The general trend was for mean IQs of children re-evaluated with the WISC to decrease slightly. Thomas explained these results in terms of test difficulty. That is, she believed the WISC-R is a more difficult test. In addition, she stated that practice effects on the WISC would be minimized by the three-year interval between testing sessions.

The Martin (1979) and Thomas (1980) studies provide some information about

the three year test-retest performance of children on intelligence tests; however, they are weakened by the fact that all the children were not re-evaluated with the same test they were initially given. Obviously using different tests introduces an unwanted source of error variance when test stability is of interest. In addition, both investigators included children with different cultural background and/or primary language in the same treatment groups without demonstrating that their performances were comparable.

Vance et al. (1981) compared the subtest and IQs obtained by a sample of 75 learning disabled and retarded children on the WISC-R after a two-year interval. Correlations between the first and second testings ranged from .53 to .80 for the six individual Verbal subtests and from .59 to .80 for the Performance subtests. As a result, Vance et al. concluded that "clinicians can be reasonable sure that subsequent prognosis for youngsters, as based on IQ [WISC-R], will not change appreciably" (1981, p. 399). Thus, the findings of this study provide some additional evidence that the WISC-R is a reliable instrument over a two year period, when used with certain types of exceptional children. Since test-retest intervals of 3 years are more typical in school districts it is of interest to know test stability over longer time intervals.

Method

Subjects

Students (N = 379) categorized as learning disabled, behaviorally impaired, mentally retarded, or unclassified, and who were re-evaluated in 1980, were selected randomly from the special education files of four large school districts (Ft. Collins. Colorado; Lincoln, Nebraska; Omaha, Nebraska; and Tempe, Arizona). As a result of this random selection, our sample of 379 ranged in age from 73 months to 240 months ($\bar{X}$ = 153.6 months, SD = 40.4 months) and was comprised of students from four racial groups: Anglo (113 males, 63 females); Mexican

American (35 males, 19 females); Black (100 males, 40 females); and Native American (7 males, 2 females). With respect to diagnostic categories, 105 (76 males, 29 females) were classified learning disabled, 12 (12 males) were classified behaviorally impaired, and 247 (157 males, 90 females) were classified mentally retarded. Fifteen students in the sample were unclassified.

Procedures

After permission was given by research committees in each school district, psychological reports and test protocols for initial and re-evaluations were examined for all students selected. Psychometric data from the Wechsler Intelligence Scale for Children-Revised (WISC-R) and the Wide Range Achievement Test (WRAT) was available for most all students for both the initial and re-evaluations. Data analyzed from the WISC-R included Full-Scale IQ, Verbal IQ, and Performance IQ. Data analyzed from the WRAT included Reading, Math, and Spelling subtest scores.

Data Analysis

Two analytical procedures, correlation and multiple regression, were employed to test the stated hypothesis. Children's initial IQ scores (FSIQ, VIQ, and PIQ) on the WISC-R were correlated with their retest IQ scores to determine the degree of association and hence stability between three-year test-retest performances. Correlations are provided for the entire sample and each racial by sex subset.

Results and Discussion

The findings of this study are based on a multi-racial sample of exceptional children and were examined to gain insights about test stability of the WISC-R and the WRAT over a three-year interval. In Table 1, the means and standard deviations are displayed for the four racial groups' test and retest performances on the WISC-R and WRAT. An examination of this table reveals at least two major

performance trends. First, average Full Scale IQs for males and females across the four racial groups did not change significantly over the three-year retest interval. Anglo males' Full Scale IQs on the average changed the least (-0.13 IQ points), while Mexican American females' changed the most (-2.11 IQ points). Second, the direction of change in standard score units on the three WRAT subtests across Anglo, Black, and Mexican American children, was in all but one instance (i.e., female Mexican Americans), negative and the average amount of change was 5.7, 4.8, 6.0 standard score units for Reading, Spelling, and Math. Change in achievement scores was not calculated for the Native American children because of the small sample with complete test-retest scores.

Correlations between children's initial and retest performances on the WISC-R and the WRAT are displayed in Tables 2 and 3. In general, the correlations in Table 2 illustrate the WISC-R is quite stable for a multi-racial group of exceptional children. Specifically, the three-year test-retest reliabilities for the WISC-R IQs (FSIQ = .85, VIQ = .82, PIQ = .78) compare well with the WISC-R standardization sample (FSIQ = .95, VIQ = .93, PIQ = .89). However, an examination of the correlations in Table 3 indicates considerable differences in degree of association between initial and retest performances on the WISC-R across the four racial groups. Clearly, the Anglo groups' test-retest correlations (FSIQ = .90, VIQ = .89, PIQ = .83) were highest and most similar to those of the WISC-R standardization sample which was comprised of 85% white and 15% non-white children (Wechsler, 1974, p. 21). Thus, our IQ stability hypothesis was supported since Anglos' IQ test-retest correlations were higher on all three IQ subscales than those of Black, Mexican American, or Native American children. It is noteworthy, however, that the stability coefficients were vary similar in magnitude to the one previous study in which test-retest reliability was examined for a group of Mexican American children (Henggeler & Tavormina, 1979).

The WISC-R test-retest stability for children in different diagnostic categories was not analyzed because of the variability in IQ cut-off scores utilized in the three states from which our sample was selected. Instead, we divided the sample into three IQ groups: less than 70 (N = 143); between 70 and 85 (N = 137); greater than 85 (N = 96). Correlations were computed between their initial and three-year retest performances on the WISC-R and are illustrated in Table 4. In general, the highest IQ group's performance was the most stable.

Regression analyses were completed for the entire sample, as well as each separate racial group to determine the best predictors of future achievement (as measured by performance on the WRAT subtests). The results indicated the three best predictors of WRAT Reading three-year retest performance were initial WRAT Reading (20% of variance), initial WRAT Arithmetic (7% of variance), and the WISC-R Information subtest (2 % of variance). The three best predictors of WRAT Spelling three-year retest performance were intitial WRAT Spelling (32% of variance), initial WRAT Reading (11% of variance), and the WISC-R Similarities subtest (4% of variance). Finally, the three best predictors of WRAT Arithmetic three-year retest performance were initial WRAT Arithmetic (22% of variance), WISC-R Verbal IQ (5% of variance), and initial WRAT Reading (2% of variance). Similar findings characterized the regressions for WRAT subtests across each racial group.

Three major implications of these data may be advanced. First, we believe it may not be necessary to routinely readminister the WISC-R to Anglo exceptional children every three-years (typical re-evalaution procedure) unless there has been documented indications of behaviors or physical illness that would question the validity of initial test results. A short-form WISC-R would seem to serve as a clinically sound alternative when retesting of Anglo children is required. This would save both the psychologist and the child time without

jeopardizing a diagnostic decision. Second, based on the relative low stability of WISC-R test-retest performances of Black, Mexican American, and Native American children in our sample, it seems advisable to continue to re-evaluate such children's IQ periodically. At the same time the low and negative correlations between WISC-R and WRAT scores for Mexican American children suggest that IQ, as measured by the WISC-R may be of minimal value in predicting academic achievement several years beyond the time of initial testing. Third, given the rather modest stability of children's performances on the WRAT over a three-year interval, it seems advisable to continue to formally re-evaluate exceptional children's academic growth.

Future investigations of IQ test stability are suggested in order to determine the influence of variables such as socioeconomic status and linguistic background. In addition, IQ test stability of children with above average intelligence is also needed.

References

Bayley, N. Consistency and variability in the gorwth of intelligence from birth to eighteen years. Journal of General Psychology, 1949, 75, 165-196.

Bloom, B. S. Stability and change in human characteristics. New York: John Wiley, 1964.

Elliott, S. N., Piersel, W. C., & Galvin, G. A. Psychological re-evaluations: A survey of practices and perceptions of school psychologists. Journal of School Psychology, 1982, in press.

Harnquist, K. Relative changes in intelligence from 13 to 18. Scandinavian Journal of Psychology, 1968, 9, 65-82.

Henggeler, S. W., & Tavormina, J. B. Stability of psychological assessment measures for children of Mexican American migrant workers. Hispanic Journal of Behavioral Sciences, 1979, 1, 263-270.

Honzik, M. P., MacFarlane, J. W., & Allen, L. The stability of mental test performance between two and eighteen years. Journal of Experimental Education, 1948, 17, 309-324.

Hopkins, D. K., & Brachs, G. Ten-year stability of verbal and nonverbal IQ scores. American Educational Research Journal, 1975, 12, 469-477.

Martin, F. Is it necessary to retest children in special education? Journal of Learning Disabilities, 1979, 12, 388-392.

Thomas, P. J. A longitudinal comparison of the WISC and WISC-R with special education pupils. Psychology in the Schools, 1980, 17, 437-441.

Vance, H. B., Blixt, S., Ellis, R., & Debell, S. Stability of the WISC-R for a Sample of Exceptional Children. Journal of Clinical Psychology, 1981, 397-399.

Zax, M., Cowen, E. L., Beach, D. R., & Rappaport, J. Longitudinal relationships among aptitude, achievement and adjustment measures of school children. The Journal of Genetic Psychology, 1972, 121, 145-154.

Table 1

Means and Standard Deviations of Anglo, Black, Mexican American, and Native American Exceptional Children's Test-Retest Performances on the WISC-R and WRAT

			WISC-R			WRAT		
			VIQ	PIQ	FIQ	Read	Spell	Math
ANGLO	Initial Test	Male	$\bar{X}$ 82.47	85.93	83.08	85.77	79.00	78.49
			SD 18.30	18.60	18.50	12.60	13.10	11.40
			N 111	111	111	95	41	95
		Female	$\bar{X}$ 70.27	76.15	70.38	75.17	73.77	75.80
			SD 12.60	16.50	14.30	9.50	8.50	9.70
			N 60	60	59	54	31	55
	Retest	Male	$\bar{X}$ 80.84	87.49	82.95	74.99	72.00	74.25
			SD 18.20	16.90	17.90	13.20	9.00	12.80
			N 106	106	109	106	57	107
		Female	$\bar{X}$ 69.18	78.04	71.26	73.56	72.43	71.66
			SD 15.80	18.60	16.70	9.90	8.90	9.60
			N 61	61	61	58	38	59
BLACK	Initial Test	Male	$\bar{X}$ 74.80	77.82	74.49	74.35	74.00	75.99
			SD 11.90	12.90	12.20	10.90	9.10	8.90
			N 97	96	99	94	76	94
		Female	$\bar{X}$ 71.90	76.00	72.55	77.62	77.28	77.86
			SD 9.30	12.20	8.80	6.20	6.80	7.80
			N 39	39	40	37	32	37
	Retest	Male	$\bar{X}$ 72.08	79.65	73.79	71.18	70.03	70.59
			SD 12.30	13.30	12.30	11.20	9.70	8.30
			N 96	96	95	97	80	98
		Female	$\bar{X}$ 68.63	80.03	72.35	73.33	71.97	72.61
			SD 9.50	12.50	9.50	8.30	7.70	7.30
			N 40	40	40	40	34	40

Table 1 (Continued)

				WISC-R			WRAT		
				VIQ	PIQ	FIQ	Read	Spell	Math
MEXICAN AMERICAN	Initial Test	Male	$\bar{X}$	79.89	87.33	80.77	80.28	77.60	79.72
			SD	12.10	13.20	12.90	12.70	10.90	10.27
			N	29	30	31	18	5	18
		Female	$\bar{X}$	72.60	83.28	75.30	79.83	78.57	77.92
			SD	10.1	13.60	10.90	11.20	9.20	9.60
			N	15	18	15	12	7	12
	Retest	Male	$\bar{X}$	78.24	90.60	82.33	70.21	67.82	69.37
			SD	14.40	14.40	12.50	10.30	16.30	8.50
			N	34	35	33	29	11	30
		Female	$\bar{X}$	70.71	82.68	73.19	75.27	80.00	71.27
			SD	12.10	13.90	11.20	9.20	9.10	5.80
			N	17	19	16	15	9	15
NATIVE AMERICAN	Initial Test	Male	$\bar{X}$	74.60	78.00	74.83	77.00	83.00	69.00
			SD	3.40	11.60	7.10	4.20	7.10	29.70
			N	5	5	6	2	2	2
		Female	$\bar{X}$	70.67	87.66	76.67	110.00	97.00	65.00
			SD	21.80	20.60	22.70	-	-	-
			N	3	3	3	1	1	1
	Retest	Male	$\bar{X}$	67.17	84.00	73.17	62.83	64.50	67.33
			SD	10.90	16.00	12.90	11.30	8.70	10.80
			N	6	6	6	6	6	6
		Female	$\bar{X}$	68.33	92.00	77.33	82.67	71.33	71.67
			SD	22.50	19.90	22.70	22.10	15.60	4.90
			N	3	3	3	3	3	3

Table 2

Correlations Between All Children's Initial and Retest Performances on the WISC-R and WRAT

			Initial Testing					
			WISC-R			WRAT		
			FSIQ	VIQ	PIQ	Read	Spell	Math
Retesting	WISC-R	FSIQ	.85	.77	.77	.20	.39	52
		VIQ	.80	.82	.64	.22	.31	.53
		PIQ	.75	.59	.78	.10	.24	.41
	WRAT	Read	.27	.34	.12	.46	.23	.33
		Spell	.17	.18	.06	.56	.57	.32
		Math	.46	.45	.35	.26	.31	.47

Table 3

Correlations Between Initial and Retest Performances on the WISC-R and WRAT for Exceptional Children From Four Racial Groups

				Initial Testing					
				WISC-R			WRAT		
				FSIQ	VIQ	PIQ	Read	Spell	Math
ANGLO	Retesting	WISC-R	FSIQ	.90	.86	.82	.18	.47	.68
			VIQ	.88	.89	.72	.20	.32	.66
			PIQ	.83	.72	.83	.09	.30	.60
		WRAT	Read	.27	.36	.05	.50	.09	.35
			Spell	.26	.28	.17	.69	.65	.44
			Math	.56	.56	.44	.26	.34	.54
BLACK	Retesting	WISC-R	FSIQ	.68	.51	.69	.38	.32	.36
			VIQ	.62	.61	.49	.46	.39	.38
			PIQ	.59	.33	.70	.20	.18	.21
		WRAT	Read	.28	.33	.26	.70	.58	.32
			Spell	.25	.29	.18	.63	.53	.29
			Math	.26	.22	24	.46	.43	.32
MEXICAN-AMERICAN	Retesting	WISC-R	FSIQ	.79	.68	.59	.42	.21	.10
			VIQ	.70	.82	.36	.35	.31	.13
			PIQ	.65	.35	.61	.30	-.01	.01
		WRAT	Read	.10	.06	-.27	.20	.70	.22
			Spell	-.18	-.51	-.55	.15	.67	.21
			Math	.24	.10	-.17	.28	.28	.52

Table 3 (Continued)

			Initial Testing					
			WISC-R			WRAT		
			FSIQ	VIQ	PIQ	Read	Spell	Math
NATIVE-AMERICAN	Retesting	WISC-R						
		FSIQ	.79	.72	.75	.95	.51	.36
		VIQ	.73	.76	.63	.84	.28	.58
		PIQ	.75	.60	.81	.53	.95	-.82
		WRAT						
		Read	.61	.57	.79	.92	.46	.42
		Spell	.68	.66	.79	.82	.25	.61
		Math	.29	.30	.36	.49	-.20	.89

Table 4

Correlations Between Three IQ Groups' WISC-R

Test-Retest Performances

				Initial Test WISC-R		
				FSIQ	VIQ	PIQ
IQ Less Than 70	Retest	WISC-R	FSIQ	.58	.38	.56
			VIQ	.52	.59	.29
			PIQ	.44	.06	.65
IQ Between 70 and 85	Retest	WISC-R	FSIQ	.51	.27	.32
			VIQ	.43	.51	.05
			PIQ	.44	-.01	.48
IQ Greater than 85	Retest	WISC-R	FSIQ	.73	.57	.49
			VIQ	.69	.70	.28
			PIQ	.55	.21	.59